Distributed Data Management in 2020?

M. Tamer Özsu[1], Patrick Valduriez[2] (moderators)

Serge Abiteboul[2], Bettina Kemme[3], Ricardo Jiménez-Péris[4], Beng Chin Ooi[5]

[1]*University of Waterloo, Canada*
Tamer.Özsu@uwaterloo.ca

[2]*INRIA, France*
firstname.lastname@inria.fr

[3]*Mc Gill University, Canada*
kemme@cs.mcgill.ca

[4]*University of Madrid, Spain*
rjimenez@fi.upm.es

[5]*National University of Singapore*
ooibc@comp.nus.edu.sg

I. Panel Overview

Work on distributed data management commenced shortly after the introduction of the relational model in the mid-1970's. 1970's and 1980's were very active periods for the development of distributed relational database technology, and claims were made that in the following ten years centralized databases will be an "antique curiosity" and most organizations will move toward distributed database managers [1]. That prediction has certainly become true, and all commercial DBMSs today are distributed.

The decade of 1990's saw the development and maturation of client-server technology and the introduction of object-orientation – both as stand-alone systems and as object-relational DBMSs.

The two moderators have been involved with this technology from the very early days. We published our book Principles of Distributed Database Systems in 1991, covering the fundamental distribution principles and techniques. The second edition was published in 1999 and included coverage of client-server systems and distributed object systems.

The third edition of the book has finally gone to print as we write this and should be out by the time of the conference [1]. During the writing of the third edition, we have been evaluating the past and contemplating the future. It has been almost twenty years since the first edition appeared, and ten years since the second edition. As one can imagine, in a fast changing area such as this, there have been significant changes in the intervening period. As we wrote the third edition, we incorporated technologies that were developed in late 1990's and in 2000's – P2P systems, data integration, database clusters, web and XML data management, stream data management, and cloud data management. It is apparent to us that the last ten years have seen an accelerated investigation of distributed data management technologies spurred by advent of high-speed networks, fast commodity hardware, very heavy parallelization of hardware, and, of course, the increasing pervasiveness of the web.

Now, the question is what is likely to happen in the next decade; or to put it differently, if there were to be a fourth edition of our book in 2020, *what would it be? what would be new?* This is the motivation for this panel.

In observing the changes that have taken place over the past twenty years of our involvement with this field, what has struck as interesting is that the fundamental principles of distributed data management still hold, and distributed data management can be characterized on three dimensions: distribution, heterogeneity and autonomy of the data sources. What has changed much since and made the problems much harder, is the scale of the dimensions: very large scale distribution (cluster, P2P, web and cloud); very high heterogeneity (web); and high autonomy (web, P2P). Also interesting to note is that the fundamental principles of database fragmentation (or partitioning), data integration, transaction management, replication and relational query processing have stood the test of time. In particular, new techniques and algorithms could be presented as extensions of earlier material, using relational concepts.

Today, to support the requirements of important data-intensive applications (e.g. social networks, web data analytics), new distributed data management techniques (e.g. MapReduce, Hadoop, SciDB, Peanut, Pig latin) are emerging and receiving much attention from the research community. Although they do well in terms of consistency-flexibility-performance trade-offs for specific applications, they seem to be ad-hoc and might hurt data interoperability.

The key questions the panel will discuss are: what are the fundamental principles behind the emerging solutions? Is there any generic architectural model, as the ones in [2] to explain those principles? Do we need new foundations to look at data distribution?

References

[1] M. Stonebraker, Readings in Database Systems. Morgan Kaufmann, 1988

[2] M. T. Özsu and P. Valduriez, *Principles of Distributed Database Systems,* third edition, Springer, 2011.